Contract No 1218396-09-F1- EI ISP GB

Study on ECMWF data

Final Report

13 March 1997

Prepared by:
Rutherford Appleton Laboratory
Chilton
Didcot
Oxon. OX11 0QX
United Kingdom

Tel: +44(1235)445723
Fax: +44(1235)445848

CLRC

rep.00217

Contract No 1218396-09-F1- EI ISP GB

Study on ECMWF data

Final Report

13 March 1997

Prepared by:
Rutherford Appleton Laboratory
Chilton
Didcot
Oxon. OX11 0QX
United Kingdom

Tel: +44(1235)445723
Fax: +44(1235)445848

Table of Contents

1. GLOSSARY

BADC	British Atmospheric Data Centre
CCLRC	Council for the Central Laboratory of the Research Councils
CEO	Centre for Earth Observation
CEOS	Committee on Earth Observation Satellites
CIP	Catalogue Interoperability Protocol
CLRC	Central Laboratory of the Research Councils
EASOE	European Arctic Stratospheric Ozone Experiment
ECMWF	European Centre for Medium-Range Weather Forecasts
EWSE	European Wide Service Exchange
MARS	Meteorological Archive and Retrieval System
NCAR	National Center for Atmospheric Research
NERC	Natural Environment Research Council
NILU	Norsk Institutt for Luftforskning - (Norwegian Institute for Air Research)
RAL	Rutherford Appleton Laboratory
SESAME	Second European Stratospheric Arctic and Mid-latitude Experiment
UGAMP	Universities' Global Atmospheric Modelling Programme
URL	Uniform Resource Locator

2. ABSTRACT

This is the final report of the small study contract "Study on ECMWF data" being undertaken by CCLRC (RAL) on behalf of the CEO.

The purpose of the study is to investigate ways in which data held at the European Centre for Medium-Range Weather Forecasts (ECMWF) can be made more easily available for use by scientists, at a cost within their budget, yet in a manner that is consistent with the ECMWF's charging policy.

3. INTRODUCTION

3.1 Background

There is a clear perception in the scientific research community [1,2] that it is difficult or expensive to obtain the meteorological data needed for undertaking scientific research. This is a barrier to it being used more widely, despite the fact that national meteorological services often provide data for scientific use at 'cost of access'. One source of data that has a potentially large customer base is the ECMWF. However, access to ECMWF data is often hampered by the sheer volume of data. The present study was carried out in order to report on the current methods of accessing data from ECMWF, to ascertain more precisely the user requirements for the data and, where possible, to recommend possible actions that could be taken to increase the availability of the data to the research community.

3.2 Standard Means of Access

In order to access and use ECMWF data for research purposes, users must first obtain the permission of their national meteorological service. Having done so, there are three standard routes through which a scientist can obtain data.

1. ECMWF run a data extraction service and will accept orders for a range of data products. These include the ability to obtain subsets of the various dataset. There is a charge for 'cost of access' at a rate that depends on the volume and type of data being requested. The charge is expressed in integer multiples of 'charging units'. One 'charging unit' is approximately 110 ECU.
2. A scientist may travel to ECMWF and run the software on the ECMWF computers that extracts the data they require. If the intended data extraction is significant and requires more than a minimal amount of computing resources (e.g. a large volume or data extraction over a prolonged period of time), an application for 'special project time' is possible. If this option is used, ECMWF charge only media costs.
3. The scientist may approach their national meteorological service and ask them to extract the data on their behalf. In this case, an appropriate charge is agreed between the national service and the scientist.

The first method is the simplest and provides the scientist with a customised dataset specifically for his/her purposes. It is reasonably inexpensive if a small amount of data is needed. However, as the charge increases with data volume, large datasets, e.g. more than a few months of global data, are correspondingly more expensive and are often beyond the reach of an average research budget. This is particularly acute

for numerical modellers and those studying global scale phenomena, who require months or even years of global data. One year's worth of 2.5 degree gridded data costs approximately 4 kECU. This problem has become particularly noticeable due to the recent availability of data from the ECMWF Re-Analysis (ERA) project. This project produced a coherent global model dataset of 15 years duration (1979-93) that is invaluable, for example, for research on climate trends and interannual variability. Surface and upper air data in spectral[1] form on model levels for the entire period amount to some 70 GB of data and would cost approximately 70 kECU to purchase via this route. The ERA project has other data products that are even more bulky.

The second method saves on direct monetary costs, but requires investment of effort on the part of the scientist. They must travel to ECMWF, learn how to run the software and then operate the programs. It is likely to be the only realistic current method of access to large amounts of data. This route, however, is likely to result in duplication of effort by scientists who require similar datasets, e.g. the full ERA dataset.

The third method opens the possibility of a scientist obtaining data at zero or small cost to themselves. However, user requirements' studies have indicated [1, 2] that this route to ECMWF data is not commonly used, possibly because the option is not widely known or because national services are not in a position to offer this service at a sufficiently low cost for the majority of researchers. There are a few exceptions to this, however, which are described below.

In the UK, the Meteorological Office has recently signed a Data Agreement with NERC, in which they have delegated to the British Atmospheric Data Centre (BADC) the responsibility for distribution of certain meteorological datasets to the UK research community. This provides effective distribution of ECMWF data within the UK through method 3. Data for the period 1979 to the present (including the ERA dataset) have been acquired and are available via ftp or Exabyte tape, free of charge to UK researchers, apart from media costs. The cost of developing this service has been funded by NERC. Initial start-up costs have been approximately £90K, with annual running costs of approximately £20K. However, ECMWF Council rules are such that distribution outside the country (UK, in this example) is forbidden and the BADC service cannot be extended to cover European distribution.

In a similar initiative, the Max Planck Institute in Hamburg, Germany is in the process of acquiring the full ERA dataset. We are not aware of the precise means by which they intend to distribute the data or whether they will make a charge. In another example, ECMWF data have been acquired by the NILU datacentre in Norway for ftp distribution to the EASOE and SESAME programmes. Interestingly, this is an example of data distribution throughout Europe, i.e. it is not restricted to a specific country. The main difference is that the data are intended for use on a specific project and are only available to that restricted user community and for the duration of the project, i.e. they are not for general dissemination to the wider research community.

[1] Terms that pertain specifically to model data are discussed in appendix C.

3.3 Summary of Current Situation

The ECMWF datasets are an extremely valuable resource of particular interest to the research community. They are not as easily accessible as they could be, primarily due to the large data volumes involved and the access costs (either monetary or staff effort). The problem is particularly acute for users who require large amounts of data, e.g. months or even years. The ease of access to the data varies across Europe. The data are relatively easy to access by UK researchers due to a recent UKMO Data Agreement. A similar initiative in Germany is likely to improve future access in that countries. However, the improved access achieved in those countries cannot at present be extended throughout Europe. This would require (a) a change in ECMWF Council rules to allow distribution of data across national boundaries, (b) explicit permission from individual national meteorological services for data distribution across national boundaries, i.e. we believe that in theory it would be possible for a data centre in country X to provide data to a user in country Y if the meteorological service in country Y gave permission (this route has not been tested and does not appear to us to be the logical way to proceed), or (c) an initiative by ECMWF themselves in which the cost of access to large datasets is reduced.

A new, alternative means of access to data at ECMWF has recently been developed which will improve access to relatively small amounts of data. ECMWF have developed a client-server means of accessing data [3]. This will allow a scientist in a member state to run the MARS client software on their own computer to extract data from the ECMWF database. Although this service is not yet generally available, it is likely to become the dominant method of obtaining small amounts (a few megabytes) of data in the future. This will be a welcomed addition. However, this method will not be appropriate for requests for large amounts of data due to the limited capacity of the network connections involved. Even taking into account the expected improvement in network bandwidth, it is unlikely that it will be feasible to transfer a 150 GB dataset such as the ERA via an electronic network within the next 5 to 10 years.

4. STUDY RATIONALE

In our opinion, the most fruitful approach to facilitate easier access and distribution of ECMWF data to European researchers is to work together with ECMWF to help to identify pre-defined off-the-shelf additional datasets that can be provided at reduced costs to the individual researchers. A prime concern is that any initiative taken by the CEO should not adversely affect the current data services offered by ECMWF. Revenue from current ECMWF data services, e.g. data extraction and data subsetting, provides funds not only for the staff to carry out those functions, but also for an excellent user support service to answer the many queries from users of the data. It is essential that this is maintained. However, if data costs are reduced, for example by offering standard off-the-shelf datasets at low cost, then revenue is likely to be maintained by the increased demand.

With this in mind, a survey of the research community was carried out in order to determine the most popular and appropriate candidate off-the-shelf datasets. The

approach taken was to design and distribute a questionnaire to investigate the requirements for meteorological data in general, rather than ECMWF data explicitly. For this reason, we included radiosonde data in our survey and also gave the user the opportunity to make his/her own suggestions for datasets.

5. THE QUESTIONNAIRE

In order to determine the demand for different types of meteorological data products, we defined nine products that we believe to be valuable to scientists working in a range of areas. We believe that improvement in the ease of access to these datasets would be welcomed by the research community and would stimulate and enhance many areas of research. In order to ascertain the level of interest and to prioritise the value of the nine products a questionnaire was developed and distributed to the research community. The text of the questionnaire is reproduced in appendix A. In addition, a tenth option was provided in which the user was asked to specify any other dataset that they would like to access which we had not included.

In order to be as specific as possible, the majority of the nine defined products are based on examples of datasets available from ECMWF. For this reason, ECMWF were consulted before the questionnaire was distributed. This was done both through a meeting with Dr. David Burridge (Director) and Dr. Massimo Capaldo (Head, Operations Department) and follow-up e-mails in which the precise wording of the text was discussed. ECMWF were positive towards the idea of the proposed questionnaire and stated that they would like to receive a report of the results of the survey and that they would be happy to receive and consider the development of further products or services that might be suggested as a result of the current study.

Dr Geir Braathen of the NILU data centre in Norway was also consulted during the definition of the data products.

5.1 Distribution of the Questionnaire

The questionnaire was distributed by electronic means. It was sent to the distribution lists for the European SESAME programme, the UK Universities Global Atmospheric Modelling Programme (UGAMP) and registered users of the British Atmospheric Data Centre (BADC). It was also placed in the 'Announcements' section of the EWSE.

Each of these was chosen to cover a different group of users. The SESAME programme involves primarily experimental scientists over all of Europe involved in studies of atmospheric ozone depletion. Their requirements for data are modest in terms of volume. On the other hand, the UGAMP programme requires large amounts of data, more than can easily be stored on an individual workstation. Users of the BADC have a wide range of requirements as individuals, as well as being members of different projects. All of these groups are primarily interested in atmospheric science. We also wished to gauge the interest from the widest possible range of scientists and we hoped to address this by placing the questionnaire in the EWSE.

5.2 Responses to the Questionnaire

We received 70 replies to the questionnaire of which 2 were rejected as they came from outside Europe.

Of the 70 replies to the questionnaire, 57 were direct replies to e-mail messages. The remaining 13 responses could not unambiguously be distinguished between replies to an e-mail message or a response to the announcement on the EWSE.

Of the 68 replies analysed, approximately half were from the UK, with the remainder from Germany (12), France (5), Italy (4), Netherlands (2), Belgium (1), Finland (1), Greece (1), Iceland (1), Norway (1), Russia (1), Spain (1), Sweden (1) and Switzerland (1).

Several responses came from groups, rather than individuals. These cases have been treated as though they came from individuals and have not been weighted in any way, since it was usually not clear how many scientists were represented by each of the replies.

6. METEOROLOGICAL PRODUCTS CONSIDERED IN THE STUDY

6.1 Description of the Products

The nine datasets defined in the questionnaire are listed below.

1. **Radiosonde Ascents**
   6-hourly height profiles of temperature, wind speed and direction, and humidity. These are available from a world-wide network of measuring sites.
2. **Upper air analyses in spectral form**
   6-hourly T106 spectral data (approximately 1.125 degrees in latitude and longitude), temperature, vorticity, divergence, vertical velocity, surface pressure (natural logarithm of), specific humidity.
3. **Upper air analyses in gridded form**
   6-hourly temperature, wind speed and direction, geopotential and humidity. These are provided on a 2.5 degree latitude and longitude grid with approximately 17 pressure levels between the ground and 10 millibars.
4. **Surface parameters**
   Temperature and soil wetness at four soil levels, snow depth, mean sea level pressure, wind speed and direction at a height of 10 metres, temperature and dewpoint at a height of 2 metres, surface roughness, albedo, skin reservoir content, percentage of vegetation, apparent surface humidity, surface roughness length for heat (logarithm of), skin temperature. These are available at 6-hourly intervals on a Gaussian N80 grid.
5. **Radiation parameters**
   Surface sensible heat flux, surface latent, heat flux, total cloud cover, surface solar radiation, surface thermal radiation, top of atmosphere solar radiation, top of atmosphere thermal radiation.
6. **Ocean-atmosphere couple modelling dataset**
   Large scale precipitation (including snowfall), convective precipitation (including convective snowfall), surface sensible heat flux, surface latent heat flux, total cloud cover, surface solar radiation, surface thermal radiation, top of atmosphere solar

radiation, top of atmosphere thermal radiation, wind stress (two horizontal components), latitudinal gravity wave stress, meridional gravity wave stress. These are available on a Gaussian N80 grid, once per day.

7. **Monthly averaged dataset**
   Averages, variances and covariances of analyses for temperature, wind, humidity, etc.
8. **Isentropic analyses**
   For example, daily potential vorticity on theta surfaces.
9. **Trajectories**
   Routine 5-day back trajectories from pre-defined positions and pressure levels every day. The output data would be latitude, longitude, pressure and temperature.

6.2 Inter-relationship between the data products

The majority of the above nine products defined above are meteorological data from computer models. The standard 'analyses' fields include parameters that describe the basic temperature and dynamical structure of the atmosphere. The dynamical structure can be in the form of wind components or, as in the basic ECMWF model, in terms of vorticity and divergence.

In order to carry out forecasting, observational data are 'assimilated' into a computer model of the atmosphere in order to produce the basic analyses. The observations come from a variety of different instruments, including for example surface observations, upper air radiosonde balloon data and satellite temperature observations. These data are generally assimilated at 6 hourly intervals. The resulting analyses are therefore a combination of observations and model data and are the best representation of the current state of the atmosphere. The analysis fields are then used as initial conditions for a forecast run of the computer models. In the case of the ECMWF model, which is global in nature, these analysis fields are available on a global basis.

The ECMWF model is a spectral model, in which the prime fields (vorticity, divergence, temperature, etc.) are held as a set of spectral coefficients. A spectral coefficient is the amplitude of a spherical harmonic function (Legendre polynomial) used to represent the field.

The dataset comprising the upper air analyses in spectral form (dataset 2) on model levels is therefore the model data in its 'raw' form. Although this represents the data in its purest form, it is not of the easiest format to use.

Many users prefer to use a dataset derived from the spectral dataset, in which the wind fields have been calculated from vorticity and divergence and in which the data have been transformed from spectral to spatial form (i.e. placed on a grid spaced roughly equally in latitude and longitude) and interpolated from model levels to standard pressure surfaces. For this reason, we have included dataset 3, the upper air analyses in gridded form. The 2.5 degree horizontal resolution chosen for this option is the same as one of the standard products available from ECMWF.

The monthly averaged dataset (dataset 7) comprise the monthly averages of dataset 3. The isentropic analyses (dataset 8) consist of potential vorticity fields that are derived from the standard analyses; they can therefore be derived from either datasets 2 or 3; derivations from dataset 2 provide the more accurate derivation.

As a by-product of the computer model forecast, the model also computes a number of fields in addition to the standard analyses. Examples include the surface parameters such as temperature and soil wetness, heat fluxes and parameters useful for ocean atmosphere coupling studies such as wind stress. The surface parameters (dataset 4) have been isolated into a separate dataset because this is likely to be of interest to researchers into land-use, hydrology, geographers and others. Somewhat more specialist subsets have also been included (datasets 5 and 6) with the radiation and ocean-atmosphere coupling research community in mind. There is significant overlap between these.

In addition to the model data we have also included a subset of the initial raw data that are fed into the productions of the analyses, namely data from radiosonde balloon ascents. These provide the most accurate measurement of the conditions locally to the observation site and are therefore particularly useful for regional studies. The dataset consists of a world-wide network of radiosonde station ascents every 6 hours.

Finally, a trajectory option has been included (dataset 9). These are highly derived datasets, in which a parcel of air is moved about with the flow of the atmosphere in order to determine where a specified parcel of air has travelled over a number of days. This dataset requires upper air wind data and hence can be derived from either dataset 2 or 3.

7. ANALYSIS OF RESPONSES TO THE QUESTIONNAIRE

7.1 Scientific Interests

The great majority of respondents indicated that their research area was connected with a study of the atmosphere. While the e-mail lists that we used for distribution were biased in this direction, the announcement on the EWSE was not. We believe that this shows that we received few, if any, responses to the questionnaire placed on the EWSE. This is relevant for two reasons. Firstly, it indicates that the EWSE may not yet be a good way to survey a group of users of Earth Observation data. Admittedly our survey was conducted over the Christmas period. However, we did receive many responses via e-mail. Secondly, it means that the results are biased in favour of products of interest to atmospheric scientists. If this bias is felt to be of concern when using the results of this study, we recommend that a fuller survey be carried out than was possible in the time available for the present study.

7.2 Table of Scores

A table showing the individual responses to the questionnaire is given in appendix B. This shows that the most popular products are, in order of priority : the upper air analyses in gridded form (dataset 3), the radiosonde ascents (dataset 1), the isentropic analyses (dataset 8), trajectories (dataset 9), upper air analyses in spectral form

(dataset 2) and the monthly averaged dataset (dataset 7). These datasets are considered individually in more detail in the next section, including recommendations on how to make them more easily accessible by the research community.

Not surprisingly, the more specialist datasets, i.e. the radiation parameters (dataset 5) and ocean - atmosphere coupled modelling dataset (dataset 6) were less popular than the more general datasets such as the analyses.

The surface parameters dataset (dataset 4) was also not rated particularly highly. This is almost certainly because of a bias in the user community that was polled. In particular, the SESAME user community is primarily directed to research of the upper atmosphere and the UGAMP community, being primarily concerned with global modelling, is less likely to require detailed information of the surface characteristics. In our experience, this dataset is more likely to be of interest to regional scale meteorologists and to researchers in the non-atmospheric research communities. The latter represent a potentially very large set of users, including those studying land-use, hydrology, terrestrial ecology, geography, geology, biology and architecture. For this reason, we recommend that this dataset should not be dismissed at this stage. We believe that a user requirements questionnaire of those research communities would discover significant interest. We note that a thorough survey of those users was outside the scope of the current study and would almost certainly require an amended questionnaire written in less technical terms, e.g. to explain the meaning of terms such as spectral, isentropic and potential vorticity.

7.3 Individual Comments from respondents

In addition to the specific replies to the questionnaires, a number of respondents made some general comments that were very encouraging of the initiative being taken. These confirmed to us the conclusions from earlier studies that there is a difficulty in obtaining data at costs within the budget of the average researcher and that increased availability of meteorological data would improve research capabilities and stimulate additional research. A selection of these comments is reproduced below.

- 'This questionnaire is a real good idea'.
- 'Thanks very much. You're doing a great job'.
- 'A 2 days organised tour of the ECMWF facilities, with hands-on exercises, would be extremely valuable for young (and mature scientists) involved in officially sponsored national or European projects'.
- 'I've only a set of seasonal average datasets at the moment. I would love a dataset of the monthly averages of the ERA data'.
- 'You have notes that I answer yes in all the boxes relative to the data access difficulty. We access the data through our National Meteo Service (Military) and this is a long story. I guess the European countries should make available to all members of ECMWF historical or more recent data on internet at a minimum charge'.
- Referring to radiosonde data: 'These data are like gold dust. Those that we have had access to, we use extensively'.

7.4 Suggested additional datasets

A tenth data option was provided in the questionnaire so that users could suggest additional meteorological datasets that had been overlooked by us. This option was taken up by very few respondents, suggesting that the nine datasets identified by us had adequately identified all major requirements. The few respondents who made suggestions included users wishing to acquire high resolution sea surface temperatures (3 suggestions) and a small number who made suggestions for more detailed products such as additional information about cloud height.

8. ESTIMATE OF SIZE OF USER COMMUNITY

The CEO requested that we provide an estimate of the number of potential users for each of the proposed datasets. These numbers are provided in Table 1. In the recommendations for further developments, we have collected datasets 4, 5 and 6 together, so we do the same here.

Table 1 - Estimated Numbers of Users

Dataset	Description	Estimated Users
1	Radiosonde	> 500
2	Spectral Upper Air	~ 200
3	Gridded Upper Air	> 1000
4	Surface Parameters	~ 1000
5	Radiation Parameters	
6	Ocean-atmosphere	
7	Monthly averages	~ 500
8	Isentropic Analyses	~ 200
9	Trajectories	> 500

In Table 2, we give the ranking of the proposed products based on the estimated number of users, the priority assigned as a result of the survey and an overall ranking.

Table 2 - Priority of Products

Dataset	Description	Ranking (No. users)	Ranking from quest.	Overall ranking
1	Radiosonde	3	2	3
2	Spectral Upper Air	6	5	7
3	Gridded Upper Air	1	1	1
4	Surface Parameters	2	7	2
5	Radiation Parameters			
6	Ocean-atmosphere			
7	Monthly averages	5	6	5
8	Isentropic Analyses	6	3	6
9	Trajectories	3	4	4

Based on advice from the CEO, the final ranking takes greater account of the number of potential users of the products.

9. RECOMMENDATIONS FOR FURTHER DEVELOPMENT

In the following sections, we provide further analysis and recommendations for each of the most favoured datasets, in order of the priority indicated by the questionnaire results. A summary of the recommendations and proposed actions for the CEO is given in section **Error! Reference source not found.**.

9.1 Upper air analyses in gridded form

This dataset contains temperature, winds, geopotential and humidity on standard pressure surfaces at approximately 2.5 degree horizontal resolution. An analysis was carried out to determine the size of this dataset, assuming the ECMWF model as the source of the data. The full analysis is provided in appendix D. The estimated volume of this dataset is 4.3 GB per year and would be a candidate for distribution, for example, on one Exabyte tape per year. Note that the surface parameters have also been included in this volume estimate, since they complement the upper air datasets and are small in volume.

Recommendation 1a: *The development of a subset of ECMWF data, to include the standard gridded analyses (dataset 3) at 2.5 degree resolution on a global basis, for distribution on Exabyte tape (1 tape per year). The 15 years of the Re-Analysis (ERA) dataset for the period 1979 - 93 could be held on a set of 15 tapes. Additional tapes for subsequent years could be produced and updated on an annual basis.*

Possible Action: *At present, this dataset cannot be distributed throughout Europe except by ECMWF themselves or by the national meteorological centres (or their agents as in the case of the BADC for UK distribution only). Details of the above recommendation have been supplied by us to ECMWF for their consideration as a possible future ECMWF off-the-shelf low cost data product. The production of this dataset for distribution needs to be an ECMWF initiative, due to the restrictions on data distribution across country boundaries. The data subset would need to be extracted, gridded and written to tape, and distributed with accompanying documentation and read software. BADC have the capability and the raw data required to produce this dataset and would be happy to collaborate with ECMWF in its production, if appropriate.*

Although an adequate survey of user requirements of non-atmospheric Earth Observation researchers was not achieved in this study, our knowledge of the requirements of those communities leads us to believe that this dataset would also be a high priority dataset for those communities. It is possible that those communities would appreciate a further subset of the data, for example covering only the European region but at higher resolution. The ECMWF model is able to provide data at greater that 2.5 degree resolution. ECMWF have increased resolution in the past few years and a horizontal resolution of approximately 0.5 degree is now available. Although this increases data volumes substantially, there would be sufficient trade-off due to the reduced area coverage to keep the dataset to a manageable size.

Recommendation 1b: *Subsequent development of an equivalent dataset at higher resolution should be considered, with European rather than global coverage.*

Possible action: *The same constraints on data distribution apply as for the 2.5 degree gridded data and hence this initiative needs to come from ECMWF. The user requirements for this high resolution data subset are likely to become more apparent once the 2.5 degree dataset is available and being used by the community.*

9.2 Surface, radiation and ocean-atmosphere coupling datasets

As already mentioned in section 7.2, we believe that a bias towards the global scale and upper atmosphere research community in the user survey has erroneously suggested that surface parameters are not of particular interest to the research community. We believe that the surface parameters (dataset 4), including temperature, soil wetness, mean sea level pressure, boundary layer wind speeds, etc., will be of immense interest to researchers in the fields of Earth Observation, land-use, terrestrial ecology, agriculture, etc. This dataset is relatively small (approximately 7 MB per day, 2.5 GB per year). The full Re-Analysis dataset could be distributed on 8 Exabyte tapes.

Additionally, since the radiation and ocean-atmosphere coupling datasets (datasets 5 and 6) are relatively small and have substantial overlap with each other, ECMWF may wish to merge these datasets to produce one comprehensive surface and radiation dataset.

Earlier we noted that it would require a more comprehensive survey of user requirements to accurately gauge the demand for this product. However, if other off-the-shelf products are to be developed, it is likely to be more cost effective to add this dataset to that project, otherwise the cost of the survey may exceed the cost of generating the dataset.

Recommendation 2: *The development of a subset of ECMWF data, to include the surface and radiation parameters detailed in section 6.1 at 2.5 degree resolution on a global basis for distribution. The 15 years of the Re-Analysis (ERA) dataset for the period 1979 - 93 could be held on a set of 8 tapes. Additional tapes for subsequent years could be produced on an annual basis.*

Possible Action: *At present, this dataset cannot be distributed throughout Europe except by ECMWF themselves or by the national meteorological centres (or their agents as in the case of the BADC for UK distribution only). The data subset would need to be extracted and written to tape, and distributed with accompanying documentation and read software. The production of this dataset for distribution needs to be an ECMWF initiative, due to restrictions on data distribution across national boundaries. BADC have the capability and the data required to produce this dataset and would be happy to collaborate with ECMWF in its production, if appropriate.*

9.3 Radiosonde Ascents

These data are direct observations and are not model products. Hence they are not specific to ECMWF. They are held by all national Meteorological Offices since they are important input data to meteorological forecasts. The data are received every 6 hours and are used operationally in real time and are then archived. In general, researchers do not require the data in near-real time. It is possible for researchers to approach their national Meteorological Office for the data, although the questionnaire results suggest that this route is not always easy or inexpensive. These data are also available from other institutes, e.g. National Center for Atmospheric Research (NCAR), USA, although there is an emphasis on US data. Our experience is that there is a general lack of information on the various sources of these data.

Most respondents stated a need for both European and for world-wide data. Approximate data volumes can be calculated. There are approximately 165 European stations and a further 720 stations world wide. Although data quantities vary between stations, an average volume is 4 MB per year per station. This would produce 660 MB per year for the European stations and an additional 3 GB per year for the rest of the world. Note, however, that as mentioned above, there are various sources of these data available elsewhere. A comprehensive survey of the availability of the world-wide data is recommended before going ahead and compiling the global dataset.

Recommendation 3: *The development of an on-line database of European radiosonde data (dataset 1) consisting of at least data from European stations, together with comprehensive information on sources of data outside Europe for those who require global coverage. Depending upon demand, this could be extended to include world-wide stations at a later date. In addition to the on-line service, the availability of historical data through the on-line database, for example covering the last 5 years, is recommended.*

Possible action: *The BADC have recently acquired this dataset for distribution to UK atmospheric researchers and this could be extended to serve the European community. The data are advertised through the WWW (http://www.badc.rl.ac.uk/) and can be extracted via ftp once an agreement has been signed by the user stating that the data will not be used for commercial purposes. The data have been acquired through a Data Agreement with the UK Meteorological Office, who have granted a world-wide licence for distribution of these data.*

Costs:

The estimated cost of extending this on-line service to the wider European community is as follows.

Workstation with sufficient memory to handle 10 simultaneous users, network connection and systems software	8 kECU
Disk space to store 10 years worth of European data	2.5 kECU
(Optional) disk space to store 10 years worth of World wide data	10 kECU
Total	**20.5 kECU**

In the first year of operation, the service would require half a year of effort to set up and operate. This would include user registration, a user support desk and

development of additional web pages describing the sources of global data. The cost would depend on where the service was sited, but might be around 40 kECU. In subsequent years, the cost would be less as it would only need maintenance and regular acquisition of new data. This might cost 10 kECU per year.

9.4 Trajectories

Trajectories are especially useful in the diagnosis of experimental measurements, since they are able to track the origin of air parcels. This service is of considerable interest to atmospheric chemists carrying out research on pollution studies, both in the stratosphere for ozone studies and in the lower atmosphere for pollution dispersion, etc. The calculation of trajectories require temperature and wind data, e.g. from ECMWF, in order to determine the position of the air parcel over time. Additional parameters, such as water vapour are also of interest.

There are a number of different trajectory codes available in Europe that have been developed for different purposes. Examples include isentropic trajectories (i.e. assuming no diabatic exchanges) developed at NILU for stratospheric ozone research and full 3-dimensional trajectories (i.e. taking into account diabatic heating) developed at Reading University and the BADC for tropospheric research.

The dates, geographical location and the pressure level at which the trajectories need to be 'launched' are very dependent on the specific research requirements. It is therefore difficult to develop a standard set of routine trajectories that will satisfy all users. Two options are possible: (a) software could be developed and provided to the user at the same time as the relevant input data (temperature, winds) and run locally on the users computer, or (b) a central service could be developed so that users may request ad hoc trajectories, perhaps filling in a form stating dates, locations etc.

The first of these options would be operationally the easiest to provide in principle, since the responsibility for running the trajectories is placed on the user. However, the data volumes that would need to be transferred and manipulated on the user's computer may be prohibitive at present. Approximate estimates of gridded data required for a standard trajectory run are 50 MB per day, giving a total of 250 MB for an average 5 day trajectory. This could be reduced by half if the service is restricted to the Northern Hemisphere only. These data volumes are still rather large for current network transfer rates. There is also the cost of supplying tested software that will run on many different types of computers. On account of these difficulties, we recommend that a central service be developed to calculate trajectories.

Recommendation *4: The development of a trajectory service (dataset 9), building upon existing services already available. Both isentropic and full three dimensional trajectories are required. The service should provide a capability of running trajectories at least in the Northern Hemisphere. Capabilities for the Southern hemisphere are desirable but not essential.*

Possible action: *A possible action is the expansion of the current isentropic service at NILU and the full 3-d service at BADC to include the wider community. We believe this would result in a substantial increase in usage of these services. Calculation of the trajectories would*

require ECMWF wind data. We would need to consult with ECMWF about whether there are restrictions on distributing these data products.

Costs:
The estimated cost of providing an on-line service to generate bespoke trajectories is as follows.

Workstation with sufficient memory to run the trajectory code efficiently, network connection and systems software	12 kECU
Disk cache for gridded data	2.5 kECU
Disk space for all gridded data (not needed at BADC)	50 kECU

The reference to a disk cache assumes that the entire gridded ECMWF re-analysis dataset (or something from which gridded data can be generated) is held in a near line form, and that only the few days worth of data needed to calculate trajectories are held directly on-line. There is also the simpler, but more expensive, option of holding all of the data on-line. If this service were to be set up at a site that already has ready access to the appropriate data (such as BADC), then the disk space for the gridded data would not be required.

The calculation of trajectories is a time consuming operation. We have costed a system that should be able to generate a trajectory within an hour. Although the time to calculate a single trajectory is long, the system can calculate several trajectories in this time if that is what a user asks for. To provide a system with a much more rapid response time would require a significantly more expensive computer.

In the first year of operation, the service would require half a year of effort to set up and operate. This cost would depend on where the service was sited, but might be around 40 kECU. In subsequent years, the cost would be less as it would only need maintenance and regular acquisition of new data. This might cost 10 kECU per year.

9.5 Monthly averaged dataset

This dataset is a condensed form of the complete ECMWF dataset, providing monthly averaged and other averaged data. It is of interest to those carrying out climate studies and as an introduction to the dataset for those who are considering the use of the full dataset. At the meeting with the ECMWF Director, we were informed that ECMWF will be compiling this dataset for the ERA period, probably for distribution on CD-ROM.

***Recommendation** 5: The compilation of a monthly averaged ECMWF Re-Analysis dataset (dataset 7).*

***Possible Action:** ECMWF are developing this dataset for distribution, probably via CD-ROM. Therefore no action is recommended for this dataset apart from the monitoring progress of its development and advertisement of its availability to the community when it becomes available.*

9.6 Isentropic Analyses

These are specialised fields, primarily required by the stratospheric ozone and upper tropospheric dynamics research communities. The high priority given to this dataset is probably a reflection of the SESAME user community, for whom such data are important.

The dataset consists of derived diagnostics such as potential vorticity on constant potential temperature (theta) surfaces. The fields may be derived from the basic analyses fields, i.e. datasets 2 or 3, although a more accurate calculation is achieved using the spectral data in dataset 3. Fields derived from the ECMWF model have been made available for distribution within specific measurement campaigns, e.g. EASOE and SESAME, via the NILU data centre. The dataset is very small and the data can be transferred over electronic networks. For this reason, an on-line service would be most effective. The provision of isentropic analyses from forecast data is also highly desirable for campaign planning, although there are likely to be commercial sensitivities here and negotiation with ECMWF / national meteorological services would be required.

***Recommendation** 6: An extension of the isentropic potential vorticity (PV) analysis service (dataset 8) provided by NILU for the EASOE and SESAME campaigns is recommended so that these data are more generally available to the wider community. The data are required on a global basis.*

***Possible Action:** The distribution of this dataset to the European research community would require negotiation with ECMWF. PV is a highly derived quantity; it is not clear to us at present whether this dataset would be bound by the same restrictions concerning distribution across national boundaries as the basic model parameters. The ability to provide isentropic analyses from forecast data for specific experimental campaigns would also be highly desirable. We understand that ECMWF Council rules have recently been amended to enable forecast data to be used by the research community. We are awaiting details of this amendment.*

9.7 Upper air analyses in spectral form

This is the most basic form of the ECMWF model analyses. The 2.5 degree gridded dataset (dataset 3) are derived from these. The data may be obtained on model levels or interpolated on to standard pressure levels. In the former case, the data are in their original state as produced by the model and are therefore as accurate a representation as possible, having been unaffected by interpolation. The spectral data are also at higher resolution than the 2.5 degree gridded dataset (dataset 3) and hence there is the possibility to interpolate onto a finer horizontal resolution. Because of their spectral representation and the data volume, these data are not straight forward to use and are generally required only by a subset of the atmospheric community.

An analysis of data volumes has been carried out and the estimated volume assuming output fields every 6 hours is approximately 13.5 GB per year (3 tapes). This volume could be further reduced to approximately 10 GB per year (2 tapes) if the data were supplied on model levels only, rather than both model levels and pressure levels. The

majority of the questionnaire respondents stated a preference for the model level data, so that the data are not degraded by interpolation.

***Recommendation** 7: The development of a subset of ECMWF data is recommended, to include the full global upper air analyses in spectral form (dataset 2), for distribution on Exabyte tape (2 tapes per year). The 15 years of the Re-Analysis (ERA) dataset for the period 1979 - 93 could be held on a set of 30 tapes. Additional tapes for subsequent years could be produced and updated on an annual basis.*

***Possible Action:** At present, this dataset cannot be distributed throughout Europe except by ECMWF themselves or by the national meteorological centres (or their agents as in the case of the BADC for UK distribution). Details of this recommendation have been supplied by us to ECMWF for their consideration as a possible future ECMWF low cost data product. The data subset would need to be extracted, gridded and written to tape, and distributed with accompanying documentation and read software. The production of this dataset for distribution needs to be an ECMWF initiative, due to the restrictions on data distribution across country boundaries. BADC have the capability and the raw data required to produce this dataset and would be happy to collaborate with ECMWF in its production, if appropriate.*

10. ECMWF METADATA AND THE CIP PROTOCOL

10.1 Description of the Work

The CEO requested that we investigate and provide an insight into the possible harmonisation of the ECMWF metadata with a standard catalogue protocol, i.e. the Catalogue Interoperability Protocol (CIP). To this end, we have obtained some metadata from the ECMWF and have put it into the form of CIP tables. In order to test the suitability of the CIP for use with ECMWF metadata, we tried to use as many of the standard CIP attributes as seemed reasonable. In this report, we present examples of guide, collection and product tables. We also note the few CIP concepts that do not match well with ECMWF metadata.

10.2 Inputs

We obtained metadata on the re-analysis products from the ECMWF from their web pages[2] and from two ECMWF documents [4, 5]. We also used the document that provides the specification of the CIP. The CEO provided us with appendix A of release B of the document [6]. However, we needed additional background information, which we obtained from the full text of release A [7]. We do not believe that the use of two versions of the document has affected the results of this work.

10.3 Caveat

The information presented in the tables is based on our understanding of the CIP from a brief reading of documents [6, 7]. We did not have sufficient time, nor did it seem necessary, to acquire an in depth understanding of the intricacies of the CIP. We believe that this has not affected the results, but it should be born in mind.

[2] http://www.ecmwf.int/

10.4 Description of Tables

There follow several tables that might form part of a description of ECMWF data. There are tables for guide, collection and product entries, as well as tables describing entities that are specific to the ECMWF.

The entries in the 'value' column fall into two categories. Values in normal type are believed to be true values, although they have not been checked for accuracy with ECMWF. Values in italics have been invented for the purpose of this exercise. They are meant to indicate that a correct value could be given if the metadata were actually made available through the CIP.

A guide descriptor for the ECMWF re-analysis project might be:

Guide Descriptor Attribute	Value
Author	*Dr. J. Bloggs*
Title	The ECMWF Re-Analysis Project
Abstract	The ECMWF re-analysis project has generated a uniform reprocessing of the ECMWF model for the period January 1979 to February 1994. Several products have been generated by this project. These are: • Upper air fields on 17 pressure levels • Upper air field on 31 model levels • Surface fields All of these are available as initialised or uninitialised analyses.
PublicationDate	*1997-01-06*
VersionId	*1.0*
GeneralKeyword	ECMWF, WEATHER, CLIMATE, GLOBAL, MODEL
Format	*Word6*
URL	*http://www.ecmwf.int/docs/era/user.doc*

The ECMWF re-analysis project has generated several products. The collection of these could be described as:

Collection Descriptor Attribute	Value
ItemDescriptorId	*cip://ciprm.ecmwf.int/123*
ItemDescriptorName	ECMWF ERA
CreationDate	1996-07-21
CollectionType	NON_TERMINAL
Purpose	This is the collection of all products generated as a result of the re-analysis project.
RevisionDate	1996-09-03
VersionId	1.0

Originator	ECMWF
General Keyword	ECMWF, WEATHER, CLIMATE, GLOBAL, MODEL
SpatialCoverage	-90, 90, -180, 180
GuideId	http://www.ecmwf.int/data/reanalysis.html
CollectionContents	*cip://ciprm.ecmwf.int/101* *cip://ciprm.ecmwf.int/102* *cip://ciprm.ecmwf.int/103* *cip://ciprm.ecmwf.int/104* *cip://ciprm.ecmwf.int/105* *cip://ciprm.ecmwf.int/106*
StorageMedium	Unitree hierarchical datastore

The list of product descriptor attributes that describes one of the ERA products might be:

Product Descriptor Attribute	Value
ItemDescriptorID	*cip://ciprm.badc.rl.ac.uk/PID.001.upi*
ItemDescriptorName	Upper atmosphere pressure level fields from the initialised analysis at T106 resolution. Results of the re-analysis project undertaken by the European Centre for Medium-Range Weather Forecasts during 1995-1996.
StartDate	1979-01-01
StartTime	00:00:00
EndDate	1994-02-28
EndTime	18:00:00
SpatialCoverage	-90, 90, -180, 180
Originator	ECMWF
MissionID	ECMWF-ERA
ArchivingCentreID	ECMWF
ProcessingCentre	ECMWF
ReprocessingActual	The entire ERA project is to perform a uniform reprocessing of data over the period of interest.
SizeMBESCProduct	220000
ModelFields	(Temperature, VerticalVelocity, Vorticity, Divergence, RelativeHumidity, Geopotential)

The product contains semantic attributes that need further definition.

Semantic Attribute	Value
NAME	ModelFields
MEANING	List of fields calculated by the model or derived from them and provided in the product.
SHORT_MEANING	List of model fields
VALUE_SYNTAX	STRING(Temperature, VerticalVelocity, Vorticity,

Divergence, RelativeHumidity, Geopotential)

Individual attributes specific to ECMWF are defined in the following tables.

Semantic Attribute	Value
NAME	Temperature
MEANING	Model calculation of the temperature of the air at a point in the model
SHORT_MEANING	Air temperature
UNITS	K
ALIAS	(T, Temp)

Semantic Attribute	Value
NAME	VerticalVelocity
MEANING	Model calculation of the upwards velocity of the air at a point in the model
SHORT_MEANING	Upwards velocity of the air
UNITS	Pa s^{-1}
ALIAS	W

Semantic Attribute	Value
NAME	RelativeVorticity
MEANING	Model calculation of the relative vorticity of the air at a point in the model. The vorticity is a measure of the local angular velocity of the air.
SHORT_MEANING	Local angular velocity
UNITS	s^{-1}
ALIAS	VO

Semantic Attribute	Value
NAME	Divergence
MEANING	Model calculation of the divergence of the air at a point in the model. The divergence is a measure of the rate at which air leaves a volume over and above the rate at which it entered it.
SHORT_MEANING	Excess rate at which air leaves a volume
UNITS	s^{-1}
ALIAS	D

Semantic Attribute	Value
NAME	RelativeHumidity

MEANING	Model calculation of the ratio of water vapour pressure to saturation water vapour pressure at a point in the model
SHORT_MEANING	The ratio of water vapour pressure to saturation water vapour pressure
UNITS	%
ALIAS	R

Semantic Attribute	Value
NAME	Geopotential
MEANING	Model calculation of geopotential at a point in the model. The geopotential is the integral of the gravity with respect to height from the surface to the grid point.
SHORT_MEANING	The integral of the gravity with respect to height.
UNITS	$m^2 s^{-2}$
ALIAS	Z

10.5 Non matching concepts

The metadata that we obtained from ECMWF generally fit well into the scheme provided by the CIP. From the list of attributes in appendix A.1 of [6] only **Frame**, **WRSPass**, **WRSScene** and **Track** appear to have no matching concept in ECMWF model data. This is not a problem as these attributes would be ignored when preparing CIP tables. A problem does occur with the attributes **MissionId**, **Sensor**, **SensorName** and **SourceName**. Model data have concepts that are analogous to these, but do not strictly fall into the current meaning of the attributes.

The meaning of **SourceName** is "The data source refers to the spacecraft, instrument, platform, ship, groundstation, or telescope, etc., that contains the sensors". It might be argued that ECMWF could be called a ground station, but this is stretching the concept a great deal. The meaning of **SensorName** is "The sensor is the instrument or hardware used to acquire the data". Here one can say the sensor is the Cray computer (a piece of hardware) used to run the model, although it is hardly a sensor.

Given the close match between the overall attribute set and those needed to describe ECMWF model data, we suggest that the meaning of these attributes be extended as follows:

MissionId:

- [Current] Unique code for the satellite/mission
- [Proposed] Unique code for the satellite, mission or computer model

SourceName

- [Current] The data source refers to the spacecraft, instrument, platform, ship, groundstation, or telescope, etc. that contains the sensors

- [Proposed] The data source refers to the spacecraft, instrument, platform, ship, groundstation, or telescope, etc. that contains the sensors, or the computer model that generated the data

Sensor:
- [Current] Information relating to the sensor, such as name as sensor modes
- [Proposed] *Not to be used for computer models*

SensorName
- [Current] The sensor is the instrument or hardware used to acquire the data
- [Proposed] *Not to be used for computer models*

If it is felt necessary to have an attribute where one can indicate the computer used to run the model, we recommend that a new attribute be defined, rather than use **SensorName**. We believe using **SensorName** would cause confusion.

10.6 Summary
In this section, we have given examples of how metadata from ECMWF could be made available through the Catalogue Interoperability Protocol. There is a fair match between the needs of ECMWF metadata and the facilities provided by the CIP. We have indicated where there are discrepancies and have made suggestions for how they might be overcome.

11. CATALOGUES FOR PROPOSED DATA PRODUCTS
We have recommended the creation of seven new data products or services. All of these involve the repackaging of existing data rather than the generation of new data. However, this does not mean that catalogues for these data presently exist in a form suitable for use via electronic networks.

The primary documentation about data from ECMWF is in the form of printed manuals[4,5]. These list all of the parameters available in a given dataset. Data from ECMWF is stored in GRIB format with parameters identified by number, so a detailed document is appropriate. Some information about the data exists on the web, but only at a high level of abstraction. An extract from one of the ECMWF's web pages is shown below.

> ERA Advanced Analysis Data
>
> Products at the resolution and on the grids of the generating system, for surface, pressure, and model levels.
>
> Analysis times 0000, 0600, 1200 & 1800 UTC.
>
> Surface - 20 parameters. (uninitialized analysis).
> Surface - 26 parameters. (initialised analysis).
> Upper air - 8 parameters (spectral - pressure levels).
> Upper air - 6 parameters (spectral - model levels).
> Upper air - 1 parameter (Gaussian - model levels - uninitialised analysis).
> Upper air - 4 parameters (Gaussian - model levels - initialised analysis).

The logarithm of the surface pressure is available as a single level field (spectral).

In order to make the metadata available in a searchable on-line catalogue, it will be necessary to prepare the data in a form suitable for that catalogue. It does not exist in such a form at present. However, since the ECMWF data is very regular, there is not a large amount of metadata to be generated. For example, the re-analysis project has generated a regular dataset with global fields every six hours for the period from 1979 to 1993. It does not seem sensible to generate catalogue records for each 6 hour global field. Rather it is sufficient to have a single entry giving the time span and details of the parameters and the type of data (gridded, spectral, surface, etc). Should a particular catalogue system require that each six-hour field have an entry, these could be generated automatically as each record would only differ in the date and time.

The radiosonde data (dataset 1) is very different in its requirements for being catalogued. There are almost a thousand sites around the world that make balloon ascents. Some of these take place four times per day, others less frequently. A given site will fail to make some ascents for a variety of reasons. In addition, the operation of a site is occasionally terminated and a new one started. This makes a catalogue of all radiosonde ascents vital if a scientist is to be able to determine where and when ascents took place. Unfortunately no such catalogue exists. The best that is available is a list of sites that make balloon ascents. An important component of setting up a database of radiosonde data would be the generation of the accompanying catalogue.

12. SUMMARY OF CONCLUSIONS, RECOMMENDATIONS AND ACTIONS

1. The current situation concerning access to ECMWF data by European researchers has been summarised (section 3.2) and a number of difficulties identified, particularly with respect to the acquisition of relatively large amounts of data. We have made several recommendations on how the current situation might be improved. In formulating our recommendations, we have taken account of the importance of not distorting the market that ECMWF currently enjoys.

2. We have suggested that the most practical solution to these difficulties would be for ECMWF to introduce the concept of cheap off-the-shelf bulk products. The products would be cheap as they would be pre-generated, not extracted on demand, and would offer no choice in the time period covered. The rationale for this suggestion is that reduced costs would stimulate the market for these products and that overall revenue would be maintained (or perhaps even increased) by an increase in demand for the product. Costs of the off-the-shelf subsets should be set at a level to recover the initial development of the subset, the production and transfer costs and also a contribution towards support of the ECMWF help desk.

3. A user survey of the research community was carried out to determine candidate datasets for these off-the-shelf products. Nine meteorological subsets of data were defined and users were asked to prioritise them to reflect their research

requirements. They were also given the chance to suggest additional datasets that may have been overlooked.

4. Analysis of the questionnaire results has shown that 2.5 degree gridded upper air analyses (temperature, winds, humidity, etc) are the data most in demand, followed by radiosonde ascents, isentropic analyses, trajectories, upper air spectral analyses and a monthly averaged dataset of ECMWF products, in that order (see section 6).

5. Specific recommendations and suggested actions have been made with regard to each of these datasets. In several, the next course of action is to pass on these recommendations to ECMWF for consideration as future standard products offered by their data service. In others, particularly the radiosonde dataset which was the second most requested dataset, there is no dependency on ECMWF and practical action to improve distribution to the European research community is possible immediately.

6. This study has concentrated on improving access to ECMWF data, since ECMWF is pan-European and is the obvious candidate for this investigation. However, the recommended actions require certain initiatives to be taken by ECMWF. If it proves impossible for ECMWF to take these initiatives, for whatever reason, we note that it may be possible to negotiate the provision of similar products from the UK Meteorological Office (UKMO). A recently signed UKMO - NERC Data Agreement provides a world wide licence for distribution of agreed meteorological data by NERC.

7. Some examples were given of how metadata from the ECMWF could be distributed through the CIP system. Places where the metadata does not mesh well with CIP concepts were pointed out and recommendations were made for rectifying them.

There follows a summary of the recommendations and actions of this study. Clearly an important issue is what ECMWF itself is intending to do.

Before undertaking this study, we approached ECMWF to outline the purpose of this study. At the same time we made a specific suggestion that the tape datasets listed in appendix D should be considered for development. Dr David Burridge, the Director of ECWMF, supported the study and requested that the results be sent to them. However, he did not commit to producing any products as a result of this survey.

Since it is not the intention of the CEO to disrupt any existing commercial arrangements for the supply of data, we suggest that the next step should be to send this report to ECMWF with a recommendation that they consider producing the low cost off-the-shelf tape products and that they distribute them through their existing mechanism at a charge that they deem appropriate. Since a large part of the existing charge relates to the cost of extracting the data from the archive, and there will be no

incremental extraction costs for the tape datasets, we expect that they can be supplied at an affordable cost.

In the following tables, the term 'a tape' refers to an Exabyte tape holding 5 GB of data. The capacity of tapes is, of course, increasing all the time.

Action 1a - Gridded (2.5 degree) Upper Air Analyses (1979-1993)

Recommendation:	Produce a set of 15 tapes containing upper air data on a 2.5 degree grid for the period of the re-analysis project (1979-1993).
Action:	No immediate action for the CEO. This recommendation has already been sent to ECMWF and the results of this study will also be sent to ECMWF. Current ECMWF rules do not permit anyone other than ECMWF to distribute this dataset across national boundaries.

Action 1b - Gridded (2.5 degree) Upper Air Analyses (1994 onwards)

Recommendation:	Produce a tape each year containing upper air data on a 2.5 degree grid for the period 1994 onwards. These data can be produced by ECMWF or an equivalent product can be produced by national meteorological services.
Action:	CEO to release an Invitation to Tender targeted at European National Meteorological services for the generation and distribution of this product.

Action 2a - Surface datasets

Recommendation:	Produce a set of 8 tapes containing surface data for the period of the re-analysis project (1979-1993).
Action:	No immediate action for the CEO. The results of the study will be sent to ECMWF. Current ECMWF rules do not permit anyone other than ECMWF to distribute this dataset across national boundaries.

Action 2b - Surface datasets

Recommendation:	Produce a tape each year containing surface data for the period 1994 onwards. These data can be produced by ECMWF or an equivalent product can be produced by national meteorological services.
Action:	CEO to release an Invitation to Tender targeted at European National Meteorological services for the generation and distribution of this product.

Action 3 - Radiosonde Database

Recommendation:	Development of an on-line database of radiosonde data and a catalogue of all the ascents.
Action:	Issue an ITT for the creation of the database and catalogue.

Action 4 - Trajectories

Recommendation:	Develop a service to provide bespoke trajectories, building on existing services.
Action:	Issue an ITT for the development of the service.

Action 5 - Monthly averaged datasets

Recommendation:	Generate CD-ROMs containing monthly averages of parameters from the re-analysis project.
Action:	No action for the CEO. ECMWF is already planning to do this.

Action 6 - Isentropic Analyses

Recommendation:	Provide an extension of the service currently offered by NILU.
Action:	Issue an ITT for the development of the service.

Action 7a - Upper air analyses in spectral form

Recommendation:	Produce a set of 45 tapes containing upper air data in spectral form for the period of the re-analysis project (1979-1993).
Action:	No immediate action for the CEO. This recommendation has already been sent to ECMWF and the results of the study will also be sent to ECMWF. Current ECMWF rules do not permit anyone other than ECMWF to distribute this dataset across national boundaries.

Action 7b - Upper air analyses in spectral form

Recommendation:	Produce 3 tapes each year containing upper air data in spectral form for the period 1994 onwards. These data can only be produced by ECMWF.
Action:	No immediate action for the CEO. This recommendation has already been sent to ECMWF and the results of the study will also be sent to ECMWF. Current ECMWF rules do not permit anyone other than ECMWF to distribute this dataset across national boundaries.

13. REFERENCES

1 CEO Pathfinder Study on User Requirement for the Global Change Atmosphere Community

2 Independent CEO Concept Design Studies: Non-Profit Research (Green) Team Final Report

3 ECMWF Newsletter number 72, (1966), publ. ECMWF

4 User Guide to ECMWF Products, publ. ECMWF

5 MARS User Guide, publ. ECMWF

6 Catalogue Interoperability Protocol (CIP) Specification - Release B, publ. CEOS Working Group on Information Systems and Services

7 Catalogue Interoperability Protocol (CIP) Specification - Release A, publ. CEOS Working Group on Information Systems and Services

Appendix A. DRAFT QUESTIONNAIRE

Dear Colleague,

RESEARCH REQUIREMENTS FOR METEOROLOGICAL DATA

We have been asked by the EC CEO (Centre for Earth Observation) to provide them with information about research requirements for meteorological data. This is an opportunity for you to say what YOU would like to have easy access to for your research purposes. Please take time out to reply to this questionnaire! You can do it simply by 'replying' to this e-mail and filling in the appropriate open brackets ().

Thank you

Dr. Lesley J. Gray
British Atmospheric Data Centre,
Rutherford Appleton Laboratory

Dr. Geir Braathen
NILU
Norway

Below, are listed ten meteorological datasets with a brief description of each one. Please would you

(a) provide your name, address, area of research and e-mail - without these your reply will not be considered valid. Any other comments you have will also be welcomed.

(b) give each of the listed datasets a rating depending on how useful you consider them to be for your research- simply fill in a number from 1-10 in the brackets. Assume (1) means most useful, (10) least useful and use (0) to indicate a dataset that is not relevant to your research. Please do not give the same rating to more than one dataset!.

(c) For the datasets that you rate highly, please answer the extra questions that are specific to that dataset.

Note - please make entries wherever you see brackets like this ().

---------- Start of questionnaire ----------

Name ()
Institute ()
Research Field (e.g. atmosphere, oceans, land surface, hydrology etc ()
e-mail ()

1. Radiosonde data : Rating = ()
---
6 hourly height profiles of temperature, winds, humidity from a worldwide network of measuring sites.

Are you interested in national, european or world-wide data? ()

Do you want historical data e.g. 1979-93? () or more recent data? ()

Do you use these data already - if so, have you had problems in acquiring them? ()

2. Upper air met. analyses IN SPECTRAL FORM : Rating = ()

e.g. from ECMWF :
6 hourly T106 spectral data (approx. 1.125deg lat x long).
Temperature, vorticity, divergence, vertical velocity,
ln(surface pressure), specific humidity between ground and 10mb.

Do you want historical data e.g. 1979-93? () or more recent data? ()

Would you prefer data on model levels or interpolated to pressure levels? ()

Is T106 resolution OK or would you need higher or lower resolution? Please give details. ()

Do you use this type of data already - if so, have you had problems in accessing / using them? ()

Are there any parameters missing from the list above - or are any surplus to your requirements?
Please give details ()

3. Upper air met. analyses IN GRIDDED FORM : Rating = ()

e.g. from ECMWF :
6 hourly temp, u, v, w, geopotential height, humidity on a 2.5 degree lat x long. grid on approx. 17 pressure levels between ground and 10mb.

Is a 2.5 degree resolution sufficient for your purposes or would you require higher resolution? ()

If so, what resolution do you require? ()

Do you want historical data e.g. 1979-93? () or more recent data? ()

If you require very high resolution, do you need global data or would a limited area e.g. Europe be sufficient? ()

Do you use this type of data already - if so, have you had problems in accessing / using them? ()

Are there any parameters missing from the list above - or are any surplus to your requirements?
Please give details ()

4. Analyses of surface parameters Rating = ()

e.g. from ECMWF :
temperature and soil wetness at soil levels 1-4,
snow depth, mean sea level pressure, 10 metre u and v,
2 metre temperature and dewpoint, surface roughness, albedo,
skin reservoir content, percentage of vegetation,
apparent surface humidity, log. surface roughness length for heat,

skin temperature.

N80 gaussian gridded data or a regular 2.5 degree lat x long grid? ()

Do you want historical data e.g. 1979-93? () or more recent data? ()

Do you use this type of data already - if so, have you had problems in accessing / using them? ()

Are there any parameters missing from the list above - or are any surplus to your requirements?
Please give details ()

5. Radiation dataset : Rating = ()

e.g. from ECMWF model :
surface sensible heat flux, surface latent heat flux, total cloud cover, surface solar radiation, surface thermal radiation, top solar radiation, top thermal radiation. Twice daily, 17 pressure levels between ground and 10mb.

N80 gaussian gridded data or a regular 2.5 degree lat x long grid? ()

Do you want historical data e.g. 1979-93? () or more recent data? ()

Do you use this type of data already - if so, have you had problems in accessing / using them? ()

Are there any parameters missing from the list above - or are any surplus to your requirements?
Please give details ()

6. Ocean-atmosphere coupled modelling dataset : Rating = ()

e.g. from ECMWF :
large scale precipitation (including snowfall), convective precipitation (including convective snowfall), surface sensible heat flux, surface latent heat flux, total cloud cover, surface solar radiation, surface thermal radiation, top solar radiation, top thermal radiation, u-stress, v-stress, latitudinal gravity wave stress, meridional gravity wave stress. Gaussian N80 gridded data, once per day.

Do you want historical data e.g. 1979-93? () or more recent data? ()

Do you use this type of data already - if so, have you had problems in accessing / using them? ()

Are there any parameters missing from the list above - or are any surplus to your requirements?
Please give details ()

7. Monthly averaged dataset. Rating = ()

e.g. from ECMWF :
averages, variances and covariances of analyses;

temp, winds, humidity etc.

Do you want historical data e.g. 1979-93? () or more recent data? ()

Do you use this type of data already - if so, have you had problems in accessing / using them? ()

Are there any parameters missing from the list above - or are any surplus to your requirements?
Please give details ()

8. Isentropic analyses : Rating = ()
---
e.g. daily PV on theta surfaces.

What resolution do you need? Is 1.125 lat x long sufficient? ()

Do you want historical data e.g. 1979-93? () or more recent data? ()

Do you use this type of data already - if so, have you had problems in accessing / using them? ()

Are there any parameters missing from the list above - or are any surplus to your requirements?
Please give details ()

9. Trajectories : Rating = ()
---
e.g. routine 5-day back trajectories from pre-defined positions and pressure levels every day. Output data would be lat, long, pressure, temp.

Do you want historical data e.g. 1979-93? () or more recent data? ()

Do you use this type of data already - if so, have you had problems in accessing / using them? ()

Are there any parameters missing from the list above - or are any surplus to your requirements?
Please give details ()

10. Your choice Rating = ()

This one is left open for you to make a suggestion that we haven't thought about!
Please give details ()

Thank you for taking the time to reply.

Appendix B. INDIVIDUAL RESULTS OF THE QUESTIONNAIRE

The table below lists the individual results from the questionnaire. It shows the score that each respondent gave to the example products. The numbers in the table have been derived from the raw numbers in the replies to the questionnaire. If the respondent rated a product with a priority of 1, it receives a score of 10, if the priority was 2, it receives a score of 9, and so on. A priority of zero receives a score of zero. Although the questionnaire stated that the products of interest should be ranked in order of priority and products not of interest should be given a priority of zero, some replies contain 'priorities' such as 1, 2, 3, 4, 8, 9, 10. We have not attempted to correct for these biases on the presumption that the respondent intended to use this to apply a weighting factor.

Product Number									
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
8	10	9	0	7	0	1	2	3	10
7	10	8	5	6	9	0	0	0	7
8	4	7	2	8	0	6	9	10	0
9	6	10	0	0	0	0	8	7	0
5	4	10	6	7	2	3	9	8	0
10	6	7	0	0	0	4	9	8	0
0	0	0	9	0	0	0	0	10	0
8	2	7	4	6	3	5	10	9	0
8	0	3	1	3	7	9	0	10	0
4	0	7	0	3	0	5	6	10	0
6	4	9	1	2	3	8	7	5	10
9	0	7	3	0	0	2	8	9	0
8	0	0	7	6	5	0	0	9	10
0	0	10	7	8	0	9	0	0	0
10	1	7	4	6	3	6	8	9	0
0	6	7	5	3	10	8	4	0	9
9	10	8	0	0	0	0	7	6	0
0	3	10	0	0	0	8	9	4	0
6	0	1	0	0	0	8	2	3	0
0	7	9	0	0	0	10	8	0	0
7	5	10	2	4	0	6	8	7	9
6	2	0	0	0	1	0	0	0	0
0	10	7	1	8	3	2	6	0	0
8	0	6	9	7	0	10	0	0	0
10	3	9	2	8	6	7	5	4	0
10	0	0	0	0	0	0	8	9	0
10	1	7	4	6	3	6	8	9	0
4	0	3	0	0	0	0	2	6	0
8	10	10	0	8	0	0	10	9	0
6	9	10	6	6	0	8	10	9	0
3	10	9	10	5	10	10	1	3	0
9	0	10	0	0	0	0	0	0	0
5	10	9	0	0	7	6	4	0	0

	6	10	0	8	9	0	0	7	0	0
	6	10	10	5	8	6	10	5	3	0
	10	8	8	10	10	8	8	10	10	10
	6	5	8	0	0	0	9	3	0	0
	5	3	1	1	5	2	1	9	9	0
	8	9	10	0	5	0	7	6	4	0
	4	9	10	0	0	6	5	8	7	0
	4	0	3	0	0	0	0	2	6	0
	1	3	3	3	6	3	1	3	1	1
	10	7	9	0	0	4	5	8	6	0
	8	4	10	0	7	0	9	10	6	0
	6	0	4	0	0	0	0	4	0	0
	8	1	9	2	4	5	3	6	10	0
	0	0	0	1	6	6	0	0	6	0
	10	4	6	8	5	0	7	0	9	0
	9	10	2	7	3	8	6	4	1	0
	4	2	2	10	0	2	4	2	4	10
	0	10	0	10	10	10	0	0	0	0
	1	0	7	6	5	3	2	8	10	9
	3	10	7	0	5	8	6	9	4	0
	9	8	10	0	8	0	9	10	9	10
	4	10	2	0	5	2	4	8	10	0
	5	0	7	0	10	9	8	6	0	0
	10	6	5	0	4	0	3	7	8	9
	9	5	10	4	0	0	0	9	8	0
	1	10	0	9	0	8	7	5	6	0
	8	6	10	0	5	0	4	9	7	0
	1	1	1	0	8	8	8	1	1	0
	9	0	10	6	8	3	4	7	5	1
	10	10	9	0	0	0	6	8	7	0
	9	0	10	8	7	6	5	0	0	0
	4	9	3	0	6	7	2	10	8	5
	0	10	8	0	7	0	0	6	9	0
	6	5	7	0	0	0	0	9	7	0
	6	6	0	0	0	0	0	1	3	0
Total	**401**	**324**	**427**	**186**	**273**	**186**	**290**	**368**	**360**	**110**

Appendix C. DESCRIPTION OF TERMS USED IN THE ECMWF MODEL

Some of the terms used in describing the ECMWF data derive from the nature of the model. The model does not represent physical fields on a normal spatial (latitude - longitude) grid, but uses a spectral representation in terms of spherical harmonic functions. This gives greater accuracy for a given computational cost. The spectral series is truncated after a certain point and the method used is called triangular truncation. Thus the ERA data is referred to as being T106 resolution, meaning that is has triangular truncation at a wave number of 106.

The ECMWF model is a three dimensional one. The vertical dimension is represented by 31 so-called model levels. In some products, the vertical dimension is interpolated on to surfaces of equal pressure.

Some products represent fields on a latitude - longitude grid. In these cases, a so-called Gaussian grid is used. This is regular in longitude, but not in latitude. Towards the poles, the lines of longitude converge, giving a dense latitude-longitude grid. However, there is no physical meaning to this increased density; it is merely a geometric artefact. The Gaussian grid uses fewer grid points towards the poles. This maintains the accuracy of the model results without using points that are unnecessarily close together.

Appendix D. ANALYSIS OF PRODUCT SIZES

Exabyte Tape Subset of Re-Analysis Dataset

Reduced resolution, 2.5 gridded data

Data are on 2.5 degree grid, standard pressure surfaces (1000, 925, 850, 775, 700, 600, 500, 400, 300, 250, 200, 150, 100, 70 ,50 ,30, 10HPa).

Data	Size per output time (00,06,12,18 hrs)
Upper air data : Geop. Ht, T, U, V, w, Spec. hum, rel. hum.	76 Mbytes
Surface fields : surf temp, soil wetness + temp, snow depth, msl press, 10m U,V, surf. roughness, albedo, % vegetation, etc (see accompanying paper for full details)	13 Mbytes
Total	**89 Mbytes**

There are 4 tar files per month, each containing the above data for a specific output time (00,06,12,18 hrs).

89 Mbytes x 4 output times = 356 Mbytes per month (approx)
= 4.3 Gbytes per year (approx)

Exabytes can hold 5 Gbytes, hence 1 year per tape.
Total Re-Analysis period = 14 tapes.

Full resolution dataset

Data	Size per output time (00,06,12,18 hrs)
spectral upper air fields (model levels) : Vo, D, w, T, LnSP	92 Mbytes
grid point upper air fields : q	69 M bytes
Grid point surface fields : surf. temp, soil temp + wetness, snow depth, sea level press, 10 metre u,v, 2 metre temp + dewpoint, surf roughness, albedo, etc (see accompanying paper for full details)	46 Mbytes
spectral upper air fields (pressure levels) : Geop Ht., T, Vo, D, w, rel. hum.	74 Mbytes
Total	**281 Mbytes**

There are 4 tar files per month, each containing the above data for a specific output time (00,06,12,18 hrs).

281 Mbytes x 4 output times = 1.1 Gbytes per month (approx)
= 13.5 Gbytes per year (approx)

Exabytes can hold 5 Gbytes, hence 4 months per tape.
Total Re-Analysis period = 14 years x 3 tapes per year = 42 tapes.

Note : it may be that users would use **either** model levels **or** standard pressure levels but not both. Hence, the dataset could be split into two separate datasets.

- Model level dataset : (92+69+46) Mbytes x 4 output times x 12 months = 10 Gbytes per year = 28 tapes total
- Pressure level dataset : (74+46) Mbytes x 4 output times x 12 months = 5.8 Gbytes per year = 17 tapes total.

(i) Each dataset would require the grid point orography, standard software to read the data and documentation files. Another possibility for consideration is the inclusion of standard software to interpolate from the spectral to grid point fields (with suitable advice on its use).

Appendix E. ADDENDUM - SIZES OF SPECIFIC DATA PRODUCTS

The CEO requested that we provide the size of three particular data products; the 2.5° gridded upper air dataset, the surface and radiation dataset and the radiosonde dataset.

Gridded Upper Air Data

The size of the 2.5° degree gridded upper air data is 3.6 GB per year, or 66 GB for the period January 1979 to December 1996.

Surface and Radiation Data

The size of the surface, radiation and ocean-atmosphere dataset is as follows:

	MB per day	GB per year
Surface data	7	2.5
Radiation parameters	9	3.3
Ocean-atmosphere	8	2.9
Total	**24**	**8.7**

Radiosonde Data

The size of the radiosonde data is 0.6 GB per year for 150 European stations. This gives 4.8 GB for the period 1990 to 1997.